Utilizing structural equation modeling to correlate biosand filter performance and occurrence of diarrhea in the village of Enseado do Aritapera in Para, Brazil
L. E. Voth-Gaeddert, D. W. Divelbiss and D. B. Oerther

ABSTRACT

Previously, our earlier work demonstrated the use of structural equation modeling to evaluate the effectiveness of point-of-use biosand filters (BSF) to reduce the occurrence of diarrhea in rural Guatemala. While prior research in laboratory and field locations has documented the effectiveness of BSF to remove agents of diarrhea, experience in field sites suggests that multiple local factors greatly influence the occurrence of diarrhea. In addition to the BSF, this study evaluated household education level, socioeconomic status, water source and handling, and sanitation as factors impacting the occurrence of diarrhea for households in the village of Enseado do Aritapera in Para, Brazil. Of the 18 correlations examined, 16 were negatively correlated, and the strongest correlation was between the utilization of an ‘improved’ water source and the reduction of the occurrence of diarrhea within the household. While proper operation and maintenance of the BSF was found to have a negative correlation with the occurrence of diarrhea, it was not the most influential factor. This result supports the prior findings from our earlier work suggesting that more research is needed to evaluate which intervention should be prioritized for maximum return on investment with aid distribution.

Key words | structural equation modeling, biosand filter, occurrence of diarrhea, Brazil

INTRODUCTION

Diarrhea is an important public health problem closely associated with drinking water and hygiene (Esrey et al. 1991; Fewtrell et al. 2005). While developed countries have created public health systems that effectively break the fecal-oral route of exposure, in developing countries the transmission of diarrhea via unclean water accounts for over 85% of the global occurrence of diarrhea (Pruess et al. 2002). More than ten million children younger than five years of age perish annually from preventable diseases (Black et al. 2005), and the occurrence of diarrhea is the second leading cause of child death in the world (Tiwari et al. 2009). Interventions that provide clean drinking water and improve hygiene are important; yet it is also important to recognize that the occurrence of diarrhea depends upon fluids (i.e., drinking water), feces (i.e., hygiene), fingers, flies, and fields (World Bank 2014) as well as the personal and collective decisions made by individuals, households, communities, and nations.

While relative risk often is used to evaluate the probability of an illness among an exposed population as compared to a population protected through an intervention, the complex etiology of diarrhea illness requires a more complex statistical tool. Structural equation modeling (SEM) captures complex relationships hypothesized among interacting variables. The construction of a SEM begins with researcher specified hypotheses. Subsequently, statistical approaches are used to evaluate the extent of causal relationships among both observed and latent variables. In a prior study, Divelbiss et al. (2013) utilized previously published research and field experience to construct
hypotheses relating the occurrence of diarrhea and the use of point-of-use biosand filters (BSF) in the Ixcan region of Guatemala.

In this study, the prior hypotheses were modified based upon field experience in Para, Brazil, and SEM was used to evaluate the occurrence of diarrhea and the use of BSF in households in the village of Enseado Do Aritapera in Para. Non-governmental agents have distributed the BSF in Para over the past decade in an effort to reduce the occurrence of diarrhea. Version 9 of the BSF designed by the Centre for Affordable Water and Sanitation Technology (CAWST; Calgary, AB, Canada) was the unit most prevalent in the field. While the BSF has proved effective in removing pathogens, parasites, turbidity and some metals (Kubare & Haarhoff 2010), the objective of this study was to use SEM to evaluate if the BSF was the primary intervention responsible for the reduction in the occurrence of diarrhea.

METHODS

Location

The village of Enseado Do Aritapera is located within the municipality of Santarem in the state of Para, Brazil on the Tapajos River (2.4500° S, 54.7200° W). Field interviews with village leaders identified that fishing is the most common form of employment for males while most females create handicraft souvenirs to sell in the market. Based upon comparison with similar villages in other countries, the research team estimated that villagers were not experiencing extreme poverty (i.e., income was greater than $1.25 per person per day). Medical facilities were reported in larger, neighboring villages, but village leaders reported difficulty in accessing the facilities.

During the rainy season, the village is flooded for approximately 3 months (March, April, and May). All houses within the village are constructed of wooden lumber, and all houses are elevated on stilts to avoid flooding. During flood season, sanitary waste is dumped directly beneath homes into the flooded Tapajos River, and drinking water is collected from the same water source. During the dry season, pit latrines are used, and drinking water is collected from the Tapajos River and stored in tanks near the homes.

Data collection

A native translator was recruited to provide assistance with the field study. During interviews with the village political leadership, homes that were considered ‘poor’, ‘average’, and ‘wealthy’ were identified. From these homes, the study team randomly selected a total of twenty households for interviews based upon the constraints of available time in the field, willingness of households to participate, and an effort to equally sample representative homes from each of the three groupings. Samples of household potable water and water supply were collected and analyzed in the field, and an interview was performed with an adult member of each household. A prior field survey was utilized (Divelbiss et al. 2013) with modifications including: translating the survey from Spanish to Portuguese, removing questions regarding floor and wall material (i.e., all homes were constructed of wood with a sheet metal roof), and adjusting questions of transportation (i.e., including travel by boat). A total of 54 questions were used, and each household was surveyed with the assistance of a translator. A complete copy of the survey is provided as a supplementary document (available online at http://www.iwaponline.com/ws/015/101.pdf).

Data analysis

To analyze the data collected from the survey, two different approaches were used, namely SEM and a correlation matrix. For this study, the software package MPLUS7, created by Muthen and Muthen, was employed to construct and analyze the SEM. A two-step approach allowed SEM to be used as an exploratory technique, namely (step 1) the analysis of acceptable latent variables was performed first, and subsequently (step 2) the analysis of fit indices for the model were determined (Divelbiss et al. 2013). Iacobucci (2010) reported that a minimum of 50 samples were required to populate an SEM model (2009); whereas, other studies have recommended larger sample sizes (Ullman 2006; Schreiber et al. 2006; Barrett 2007). Due to the small sample size of 20 households interviewed in the field, a total of 60 bootstraps were performed on the data as previously described (Grace 2006; Kline 2005). For the purposes of this study, chi-square, comparative fit index, Tucker Lewis index, and root mean square error
of approximation were used to test model fit (Hu & Bentler 1998; Bentler & Yuan 1999; Barrett 2007; Schreiber et al. 2006; Ullman 2006). Residuals for each hypothesis were analyzed and those with a significance level of $p < 0.10$ were identified as agreed.

Figure 1 presents the conceptual model previously reported by Divelbiss et al. (2013). Boxes or ovals with the names of variables are included, and the hypothetical relationships among variables are captured through directional lines. Boxes signify observable variables and ovals signify latent (i.e., unobserved) variables. Circles encompassing the letter ‘e’ represent measurement error of observable variables. The directionality of the lines designate the direction of causality. For example, as shown in Figure 1 an improved water source is hypothesized to reduce diarrheal health burden. An ‘improved’ water source can be directly observed in a village through interviews with villages (i.e., an interviewer can see a tube well with pump). Adequate sanitation is also an observable variable as an interviewer can see a pit latrine. In contrast, household education level is a latent variable, and is therefore presented as an oval.

Latent variables are assessed through measurement of three or more observable variables. For example, the interviewer can directly assess paternal education level (i.e., by asking the question, ‘for how many years did the father attend school?’), maternal education level, and the predominant household language, and these three observable variables can be used to describe the latent variable of household education level. The directionality of the lines connecting latent and observable variables is important. For example, one would expect a high household education level to positively correlate with paternal education level; hence the arrow reflects this relationship. Household education level is hypothesized to reduce diarrheal health burden, but it is also hypothesized to improve filter operation (FOM) and maintenance. Thus, while FOM and maintenance is an observable variable, the SEM allows the user to hypothesize both a direct relationship between household education level and diarrheal health burden as well as an indirect relationship with FOM and maintenance as the intermediary variable.

As shown in Figure 1, diarrheal health burden is a latent variable contained within an oval. This variable is assessed indirectly through direct measures of observable variables including: the presence of children under five years of age; persistent reports of diarrhea in the household; at least

---

**Figure 1** | Divelbiss et al.’s (2013) SEM model used in Guatemala. Boxes represent observable variables, circles represent latent variables, ‘e’ represents measurement error and arrows represent direction of hypothesized causality.
1 day with diarrhea in the past 2 weeks; and 3 or more days with diarrhea in the past 2 weeks. The directions of the arrows capture the hypothesis that an increase in diarrheal health burden would lead to an increase in each of these observable variables.

The second approach is an analysis of a correlation matrix. When working with latent variables, the correlations between the observable variables that collectively represent the latent variable need to correlate strongly (as compared to other correlations) and consistently (±0.20) (Grace 2007). To assess whether consistently representative variables are selected for a latent variable, the representative or observable variables should be within a similar range implying likeness but not being repetitive. It is possible to assess trends in correlations with other observable variables outside of the individual latent variables. This allows for analysis of all observable variables independent of indirect influences from other variables. Walsh (2009) offers guidance on defining significance levels; however, the reader and author should be cognizant of what is under assessment. Excel 2007 was utilized and the Equation (2)/n^{1/2} where n is the sample size (n = 20) was used. This means that any value that is found greater than 0.447 can be thought of as significant. An understanding of the variables under analysis is crucial for constructive feedback. The observed trends provide corroborating support for the results observed through bootstrapping. Collectively, these results can be used in aggregate to test the validity of the originally hypothesized SEM model.

**LIMITATIONS**

SEM has been described as a combination of exploratory factor analysis and multiple regression (Schreiber et al. 2006). SEM allows for the analysis of relationships between independent variables and dependent variables where either type can be represented by a measured variable (i.e., directly observed) or a latent variable (i.e., unobserved, not directly observed) (Ullman 2006). SEM has been used extensively in psychological, social, and behavioral sciences (Bentler & Yuan 1999) as a confirmatory technique or as an exploratory technique (Schreiber et al. 2006).

It is important to recognize limitations in the approach utilized. The use of SEM as an exploratory technique arises from iterative attempts to populate the theorized model. Because repeated steps are involved, as fit is evaluated, failure occurs and iteration drives the model towards a better representation of the relationships present within the environment. Within the MPLUS7 software package, residual values are provided in tabular format allowing the user to evaluate the fitness of variables. In ideal cases, reconstruction of the conceptual model can improve fit, reduce residuals, and lead to an improved understanding of the environment. Selection bias and methods used in SEM and the correlation matrix can influence the results.

Within data collection the local political leaders were recruited to assist in the identification of twenty ‘random’ homes. While the households were selected at random, it was necessary to use the assistance of the local political leader to identify the population in which to create the list of households. Hence, the potential for bias could exist within household selection.

It is important to understand some of the major limitations of the SEM approach. When used as an exploratory tool (through iterations), hypotheses are generated first and then supported or not supported by collected data. If no relationship is hypothesized, it will not be recognized by an SEM analysis. A large unexplained variance will highlight crucial hypotheses not included in the model. However, it is then the researcher’s responsibility to conduct a review to understand the missing variables. In the current study, minimal variance was found within the model.

The iterations of SEM are important in the process of assessment. The term small data set does not imply an inadequate data set. The feasibility study is a crucial step in the exploratory process. Bootstrapping is used to increase the data set for the feasibility study and as the process is moved forward incremental increases in data set size is obtained. As the size of the original data set that is used for bootstrapping increases, overestimation is possible. However, with a smaller data set this is limited.

Finally, in the analysis of the correlation matrix, the user is limited to only assessing the correlations and cannot draw conclusions of causality from the results. However, as with all the limitations acknowledged, as
long as the user is aware of the limitations, the plurality of independent techniques increases confidence in the overall result.

**RESULTS**

A graphical representation of the conceptual model used previously by Divelbiss et al. (2013) is provided in Figure 1. This graphical representation is based upon the following 12 hypotheses (it should be noted that the term ‘negative effect’ and ‘positive effect’ reference correlations as causality is only established through several iterations):

- Increased household educational level has a negative effect on the number of occurrences of diarrhea.
- Increased socioeconomic status has a negative effect on the number of occurrences of diarrhea.
- Poor hygiene practices have a positive effect on the number of occurrences of diarrhea.
- Additional water treatment beyond the filter has a negative effect on the number of occurrences of diarrhea.
- Access to an improved water source has a negative effect on the number of occurrences of diarrhea.
- Access to adequate sanitation has a negative effect on the occurrences of diarrhea.
- Proper water storage has a negative effect on the number of occurrences of diarrhea.
- Increased socioeconomic status has a positive effect on FOM and maintenance.
- Better personal hygiene practices have a positive effect on FOM and maintenance.
- Additional water treatment has a positive effect on FOM and maintenance.
- Increased household educational level has a positive effect on FOM and maintenance.
- Proper FOM and maintenance is expected to have a negative effect on the number of occurrences of diarrhea due to the results from case-control studies (Aiken et al. 2011).
- Access to an improved water source has a negative effect on FOM and maintenance.

The reason this hypothesis was established was due to the increased focus on improved water sources by foreign aid organizations (WHO 2012) and local living conditions of villagers. The World Health Organization (WHO) has been strongly supporting access to improved sources of water globally. Locally, all houses were erected on stilts, and at the time of the administration of the survey, the Tapajos River flooded each of the homes sites. Flooding limited villager’s choices in terms of a drinking water source. Using observations from the field, a modified conceptual model was hypothesized for the Brazilian environment (see supplementary material, available online at http://www.iwaponline.com/ws/015/101.pdf).

Relationships (represented by the arrows) within the model can only be removed due to lack of significance. The dynamic between the number of regressions and the number of variables being regressed dictates whether the model is identified or not. If a variable does not offer a sufficient amount of information towards the model then the software program identifies this and removes it with permission from the user. It is important that the researcher not rely solely on the statistical analysis but consult relevant literature and on ground experience before adjusting the model.

As part of the determination of socioeconomic status, the ownership of a personal boat was added to help represent the latent variable. The observable variables for the diarrheal health burden were changed slightly in hopes of allowing for a more significant correlation. Children under the age of ten were noted, and diarrhea lasting more than 1 day was used as a variable. In looking at the independent observable variables, the only change made was to the ‘Soap Present in Home’ as the use of soap was promoted by the village leaders already within the community. An additional requirement was added in the presence of a towel. These changes were made before data collection began and then reassessed after.

The hypothesized model for Brazil was modified after bootstrap analysis with the oral survey data. The final model is presented in Figure 2. Direct relationships are given coefficients, which are reported next to the corresponding arrow. The value of the coefficient is interpreted as a one-point increase in the independent variable.
produces the given values increase (or decrease) in the dependent variable. For example, as Additional Water Treatment increases by one point, Filter Operation and Maintenance will increase by 0.3 points, therefore having a moderately positive effect on Filter Operation and Maintenance. The latent variables are represented by circles and the observable variables are represented by squares. In the model, the strongest correlation between two variables is the effect the household education level (HEL) has on the Filter Operation and Maintenance (2.5 point increase). Socioeconomic status (SES) has the second largest effect amongst all relationships with a 1.0 point increase on Filter Operation and Maintenance. For the diarrheal health burden (DHB) variable, the strongest correlation is with the HEL with −0.5. A negative value in this scenario means that as the education level increases within the household the diarrheal burden decreases. Similarly with the Filter Operation and Maintenance, a −0.1 value is found to directly affect the diarrheal burden in the house. All values presented were significant at a \( p < 0.10 \) level.

Evaluation of the latent variables was also performed. The representation within HEL and SES variables only utilize two observable variables. The recommended minimum is three observable variables with similar correlation factors (Grace 2006). However, since the model is still identified this is acceptable. The latent variable DHB has three representative observable variables, however the correlation factors are not strong indicators of good representation. Aware of these concerns, for the purposes of using SEM as an exploratory technique the trends in correlations are presented in Table 1.

Table 2 shows the correlation matrix that was computed from the data utilizing Excel 2007. The highlighted squares show the relationships that are significant. Significant meaning the value is greater than \( 2/n^{1/2} \), where \( n = 20 \). This means that any value that is found to be greater than 0.447 can be thought of as significant (Walsh 2009). It can be observed that the strongest correlation was between IWS (improved water source) and diarrhea. This correlation was negative, supporting what is already known, ‘drinking contaminated water results in more diarrhea.’ The second highest correlation is between being able to afford water being pumped to the house and IWS (improved water storage) within the house. This is a positive correlation. Of the six significant correlations found in the matrix, half of them were associated with the SES of the household. Of the six SES variables assessed within the survey, all had an average negative correlation with DHB. Of the 18 total correlations between the factors associated with SES (6) and factors associated with DHB (3), 16 had negative correlations, four of which were of low to moderate significance. These are identified in Table 2 by an outlined box.

With the small data, residuals were not available, however, improvement factors to model could be assessed.
Table 1 | Results from SEM model for each hypothesis; the outcome, the effect, and correlation

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Agreed/</th>
<th>Direct/Indirect</th>
<th>Correlation factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>agreed</td>
<td>direct</td>
<td>–0.5</td>
</tr>
<tr>
<td>2</td>
<td>agreed</td>
<td>indirect</td>
<td>SES to FOM to DHB</td>
</tr>
<tr>
<td>3</td>
<td>agreed</td>
<td>indirect</td>
<td>STP to FOM to DHB</td>
</tr>
<tr>
<td>4</td>
<td>agreed</td>
<td>indirect</td>
<td>AWT to FOM to DHB</td>
</tr>
<tr>
<td>5</td>
<td>disagreed</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>6</td>
<td>disagreed</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>7</td>
<td>agreed</td>
<td>indirect</td>
<td>In FOM score</td>
</tr>
<tr>
<td>8</td>
<td>agreed</td>
<td>direct</td>
<td>1.0</td>
</tr>
<tr>
<td>9</td>
<td>agreed</td>
<td>direct</td>
<td>0.6</td>
</tr>
<tr>
<td>10</td>
<td>agreed</td>
<td>direct</td>
<td>0.3</td>
</tr>
<tr>
<td>11</td>
<td>agreed</td>
<td>direct</td>
<td>2.5</td>
</tr>
<tr>
<td>12</td>
<td>agreed</td>
<td>direct</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

Original hypotheses established before data collection.
Agreement of findings with the original hypothesized relationship at a $p < 0.10$ significance level.
Direct relationships were those found connected by one arrow, indirect relationships were those variables related through an intermediary variable.
Path coefficient representing the strength of the exogenous variable on the endogenous variable.

Table 2 | Correlation matrix from raw data

Green is latent variable HEL, orange is latent variable socioeconomic status, and brown is latent variable DH. The full color version of this table is available online at http://www.iwaponline.com/ws/toc.htm
through the correlation matrix. A pattern of consistent, moderate to high significance levels within each of the three latent variables within the model was assessed. This provides assurance that the selected indicators were appropriate for their respective latent variable. The lack of representation within the latent variable HEL is evident in only two poorly correlated observable variables. The observable variables for DHB had better correlation but still needs improvement with the possibility that an additional variable might be needed. SES had the best representation in its observable variables, however, an increase in the correlation factors between three or four variables is needed.

DISCUSSION AND CONCLUSION

SEM model

The original hypotheses are presented in Table 1 along with the results of the SEM analysis. This includes the support or failure to support of the respective hypothesis, the type of effect present (direct or indirect), and the correlation factor associated. Only two of the 12 hypotheses were either not analyzed or had no significance. Hypothesis 5 was not analyzed due to no variation among villagers reporting.

Education (HEL) was shown to have the largest effect on both the FOM and maintenance and the diarrheal health burden. These two results support their respective hypotheses, however, education was found to have a negative effect on FOM and maintenance in Divelbiss et al. 2013. As preventing sickness through consuming unclean water (i.e., fluids) is only one of the five pathways in the F-diagram to contract diarrhea, it is not surprising that a general higher education level helps in reducing other areas of potential contamination.

The BSF was shown to have a direct negative effect on the diarrheal health burden. This supported the original hypothesis and is confirmed in Divelbiss et al. 2013. This is also confirmed in the numerous reports on the BSF in field studies (Palmateer et al. 1999; Stauber et al. 2006; 2009).

While education (HEL) had the largest effect on FOM, socioeconomic status (SES) had the second highest effect on the proper use of the BSF. The higher the economic status of the household within the village, the better the care was for the BSF. However, within the SEM model it did not correlate with a significant decrease in diarrheal occurrences found within the household. The SES indicator did have issues with large variances within the bootstrapped data sets and warrants further consideration in a larger data set. This large variance produces a larger confidence interval, but does not warrant sign change (+/-) to the correlation factor and at most, minimally effects model results. Interestingly, the significance of education within the household was more important to decrease diarrheal issues within the house than the presence of the BSF. This result is consistent with Divelbiss et al. (2013), and suggests that while access to a BSF may be helpful, there are benefits to other types of interventions if the ultimate objective is to reduce the occurrence of diarrhea illness.

Correlation matrix

The correlation matrix is presented in Table 2 as discussed in the results. Caution is needed when analyzing a correlation matrix as causality cannot be determined from the data. The matrix provides more information regarding the variables associated with water and the SES relationships that were not significant in the model. SES variables were shown to have a consistent negative correlation with diarrhea variables. This suggests that either having an increased economic status allows access to better health or that better health allows for more opportunity for the advancement in economic status. Several negative correlations were found between the issue of diarrhea and variables associated with water (i.e., improved source, additional water treatment and improved storage) and sanitation (i.e., improved sanitation). UNICEF and many non-governmental organizations devote large resources to improvements in local water sources; this is reinforced in the research project as an important topic. However, the data also show the importance of a household’s socioeconomic status.

In comparison with the earlier study by Divelbiss et al. (2015) in Guatemala, similarities between results is apparent. The importance of education and its effect on the overall diarrheal health burden is reported in both projects. SES is also a primary or secondary factor in the health of the household. Sanitation was found to be a significant factor indirectly to
the reduction of diarrheal occurrences. The BSF's negative effect on diarrhea reported in subsequent articles is confirmed within both projects (Kubare & Haarhoff 2010). Water management in regards to health is important, however, these reports promote the importance of having a holistic understanding of the complex set of relationships within developing communities to aid in effective interventions. While these findings add to the literature in identifying important issues within aid, further research is needed to warrant change within the development community.

REFERENCES


Walsh, D. 2009 A Simple Rule of Thumb for Statistically Significant Correlation. Middle Tennessee State University, 1–6.


First received 10 December 2013; accepted in revised form 19 September 2014. Available online 7 October 2014